International Conference on Multimedia and Expo (ICME'04), Taipei, Taiwan, June 27-30, 2004

TOWARDS ROBUST FACE RECOGNITION FROM MULTIPLE VIEWS

Ming-yu Chen and Alexander Hauptmann

School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213

ABSTRACT

This paper presents a novel approach to aid face recognition: Using multiple views of a face, we construct a 3D model instead of directly using the 2D images for recognition. Our framework is designed for videos, which contain many instances of a target face from a sequence of slightly differing views, as opposed to a single static picture of the face.

Specifically, we reconstruct the 3D face shapes from two orthogonal views and select features based on pair-wise distances between landmark points on the model using Fisher's Linear Discriminant. While 3D face shape reconstruction is sensitive to the quality of the feature point localization, our experiments show that 3D reconstruction together with the regularized Fisher's Linear Discriminant can provide highly accurate face recognition from multiple facial views. Experiments on the Carnegie Mellon PIE (Pose, Illumination and Expressions) database containing 68 people's faces with at least 3 expressions under varying lighting conditions demonstrate vastly improved performance

1. INTRODUCTION

Face recognition is a fascinating problem in computer vision. Many important commercial applications would be enabled by robust and accurate face recognition technology, such as identity verification, criminal face recognition, and surveillance. Nowadays, more and more video information is collected and stored in multimedia archives. The human face is a prime focus for research and is also frequently an interesting topic for retrieval from multimedia content [1].

In general, there are two different approaches to face recognition. The most well known is the family of "Eigenfaces" [2] recognition algorithms while the other is feature-based recognition [3]. The Eigenfaces approach encodes the whole face using principal component analysis which captures the greatest variations in faces and constructs an eigenspace to represent the variance. Faces are then projected into this eigenspace. Feature-based recognition derives distance and position information from facial features, like eyes, nostrils and mouth, to represent the face. More advanced featured-base algorithms construct a generic graph [4] to represent a face. The graph nodes are located at well-defined facial features and the edges are labeled with distances between the nodes. Recognition is then based on the similarity of the graphs.

Both these face recognition methods are all fairly efficient and mature. However, previous work mainly focused on static images. With the increased importance of video, the question arises: How can we get more information out of a target face in a video sequence, to assist in face recognition? Experimental evidence from psychology [5] shows that video enables people to better recognize a person compared to static pictures. Thus, we have a reason to believe that spatio-temporal information can indeed help recognition. Our goal is to utilize the constraints provided by 3D models to improve recognition. We start with a feature-based approach, which finds selected facial features in each image and then reconstruct the 3D face shapes from two orthogonal views [6]. We select a subset of features based on the pair-wise distances between points on the 3D face model using Fisher's Linear Discriminant (FLD). We denote these features as our 3D facial feature vectors and finally measure similarity to other faces with a Euclidean metric.

Our experimental study is based on the Carnegie Mellon PIE [7] database, which contains 68 people's faces, each with at least 3 expressions. Compared to previous work on Eigenfaces and feature-based recognition algorithms, our approach reduces the error rate from 12% for Eigenfaces and 15% for feature-based recognition to 1%. Along the way, we also solve an inverse and instability problem of FLD. Experimental results demonstrate that regularization of FLD not only provides the best error rate for recognition, but also makes recognition more robust and resistant to errors in the test data.

2. 3D FACE RECONSTRUCTION

In this paper we want to characterize a new approach demonstrating that multiple views enhance the ability to recognize human faces. Since we want to avoid an overly complicated system, which will confound different sources of errors in the experimental evaluations and make it harder to identify the sources of accuracy, we will

base our experiments on manually extracted features. We extracted a number of facial features from the frontal view and side view of human faces. Before we can start to use these feature vectors to construct a 3D head model, there is still the problem of normalization. Because of different zooms and views, we must normalize the feature vectors to lie on the same level. Then we can divide the vertices of the generic model into two sets, feature vertices and non-feature vertices. Feature vertices correspond to the facial features that were extracted from the available images. Non-feature vertices are the remaining vertices in the generic model. The generic model is adapted to the feature vertices and through bilinear interpolation of the non-feature vertices. Head-model construction proceeds as the follows:

1. $\{FVf_1, FVf_2 \ldots\ldots FVf_n\}$ is the set of facial feature vectors extracted from the front view of the face and $\{Vf_1, Vf_2 \ldots\ldots Vf_n\}$ is the set of corresponding vectors in the generic model. $\{FVs_1, FVs_2 \ldots\ldots FVs_m\}$ is the set of facial feature vectors extracted from the side view of the face and $\{Vs_1, Vs_2 \ldots\ldots Vs_m\}$ is the set of corresponding vectors in the generic model. Because the front view and the side view model are processed as independently, we only describe the construction of the front view in the following example. Both sides are combined together at the end.
2. We define a distance vector between facial feature and the corresponding vertex as:
$$DSf_i = FVf_i - Vf_i \qquad (1)$$
i denotes the ith vertex. The distance vectors give us the information about the difference between the real head and the generic model.
3. Based on the distance vectors we calculated from the feature vertices, we need to estimate the distance vectors for the non-feature vertices bye interpolating the distances of nearby feature vertices. The estimation is based on the following equation:
$$DSf_i = \sum_{k=1}^{N\in(d_k\le d_r)} \left(\frac{d_k}{\sum_{t=1}^{N\in(d_t\le d_r)} d_t} DSf_k \right) \qquad (2)$$
where d_k denotes the distance from ith non-feature vertex to kth feature vertex in the generic model, d_r is the range around the non-feature vertex.
4. We repeat step 2 and step 3 again for the side view, and then modify the original generic model to the new individual model by the following:
$$V_i = DS_i + V_i \qquad (3)$$
where V_i is Vf_i in the front view and Vs_i in the side view and DS_i is DSf_i of the front view and DSs_i of the side view. Figure 1. shows the result of the front view and side view mesh models after interpolation.
5. Now, we have two 2D meshes for an individual person. Because we choose orthogonal views, it is very easy for us to construct a 3D model. We denote the vertex in the new 3D model as (x, y, z) and the corresponding vectors in the frontal view and side view are (x_f, y_f) and (z_s, y_s) respectively. The 3D coordinates can be estimated as follows:
$$(x, y, z) = (x_f, \frac{y_f + y_s}{2}, z_s) \qquad (4)$$
Figure 2. shows the 3D model constructed by the algorithm.

For similarity calculation in recognition, we must represent the face models as feature vectors. We have two sets of facial features extracted from faces, one is from the front view and the other is from the side view. We also know the corresponding vertices in our individual 3D models. We define the distance between every two facial feature vertices as the feature vectors of the 3D model.

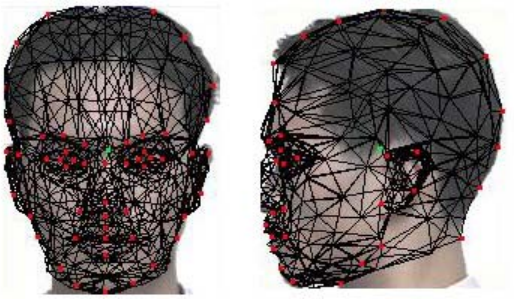

Figure 1. The front and side view of the face. The mesh is modified with the measured individual facial features (red dots) and interpolated to each individual face.

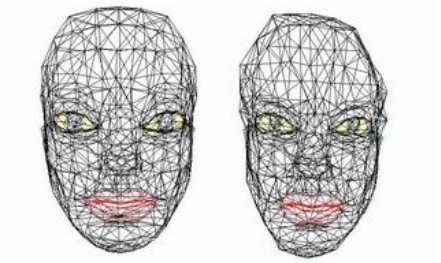

Figure 2. On the left is the generic face model while the right image contains the individual face model interpolated according to Figure 1.

3. FEATURE SELECTION BY FLD

Human faces are very similar in structure and shape. We selected the facial features to represent faces based on our intuitive knowledge and availability, but we don't really know if these features represent a discriminating set for classification. *Fisher's Linear Discriminant* is a useful feature selection method. It tries to shape the scatter and

make it more reliable for classification. The basic idea of FLD is to select *w* that maximizes the ratio of the inter-class scatter and the intra-class scatter.

The intra-class scatter matrix is defined as:

$$S_w = \sum_{i=1}^{c} \sum_{x_k \in C_i} (x_k - u_i)(x_k - u_i)^T \quad (5)$$

and the inter-class scatter matrix is defined as:

$$S_b = \sum_{i=1}^{c} |C_i|(u_i - u)(u_i - u)^T \quad (6)$$

where u_i is the mean of class C_i, u is the mean of all the data, and $|Ci|$ is the number of class Ci.

The goal is to find an optimal projection that will maximize the distance between classes and minimize the distance within the same class. Therefore, the target function to reach our goal is:

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \quad (7)$$

where w is the possible projection. The optimal projection w_{opt} is:

$$w_{opt} = \arg\max_w J(w) \quad (8)$$

From the Lagrange Multiplier Rule, we find that

$$S_b w = \lambda S_w w \quad (9)$$

This is equivalent to solving a generalized eigenvalue problem. The optimal w is the eigenvector corresponding to the largest eigenvalue of the equation 11.

An important problem in FLD is that the intra-class scatter matrix is close to a singular matrix. This is because the dimension of the feature vectors is often much larger than the number of training examples. There are two major approaches to solve this problem. The first one is to reduce the dimensionality of the feature vectors. The well-known Fisherfaces algorithm uses this approach to solve the singularity problem. It reduces the dimension by Principal Component Analysis (PCA) and then applies FLD on the reduced feature vectors. The second approach is to stabilize the intra-class scatter covariant matrix by regularization. The formulas of regularization are the following:

$$S'_w = \alpha S_w + (1 - \alpha) S_{w0} \quad (10)$$

where S_{w0} is the diagonal matrix of S_w; α is a parameter between zero and one, and is optimized experimentally.

The regularization approach not only solves the singularity problem but it also provides the ability to overcome noise in the data. In the facial feature extraction process, it is virtually impossible to get results without any error. Typically, some facial features will include noise after extraction. Because regularization assumes that features are statistically independent in the intra-class scatter matrix, some features containing errors won't affect the other features. This implies that by using FLD to reduce the dimensionality of the feature vectors, we have a better chance to reduce those non-discriminant dimensions which also contain errors and result in misleading class boundaries in the classifier.

As the experimental results show, the regularization approach provides better performance with respect to error robustness. We will discuss the experimental results in the next section.

4. EXPERIMENTAL RESULTS

Our experimental data was obtained from the CMU PIE database, which contains 232 faces from a total of 68 people. All people have at least 3 expressions, neutral, smiling and blinking (eyes closed). Some people, who wear glasses, also have another expression, neutral without glasses. All the images in PIE database are 640*486 pixels by 16777216 colors (24 bits).

We randomly chose two expressions from every person as the training set and the remaining expressions as the test set. For each experiment, we repeated the random selection 20 times and reported the average error rate as the result.

First, we want to investigate t if the spatial information is helpful. 30 facial features are manually extracted from the frontal faces and 20 facial features are manually extracted from side view. The 3D head models are reconstructed using those facial features. The distances between pair-wise facial features result in feature vectors which represent the faces. Nearest Neighbor (NN) with a Euclidean distance metric is performed to recognize faces. We used the results of frontal view eigenface recognition as a reference baseline. The eigenface method is implemented by normalizing faces to 64 by 64 pixels and reducing them to 50 dimensions using PCA. The synthesis of the frontal feature vector together with the side feature vector will provide a better comparison with the 3D approach since information from both views is exploited in either case. Figure 3 shows that there is a dramatic improvement from 2D synthesis to 3D approach. The major difference between the 2D synthesis and 3D approach is that the distances from both views are only represented in the plane but are not used to provide spatial information.

Next, we wanted to investigate the issue of feature selection. As discussed in the previous section, there are two approaches to solve the singularity problem of the intra-class scatter matrix. One is to use PCA to reduce the dimension of the feature vectors and the other is to add regularization into the intra-class scatter matrix. As Figure 3 shows, feature selection gives an additional improvement from an error rate of 7.45% to 3.27% or better. It also shows a significant improvement for the regularization method to solve the singularity problem compared to the PCA approach.

Although we extracted the facial features manually, we wanted to discover the noise resistance for both approaches. We added Gaussian noise with different standard derivations. The results in Figure 4 show that the regularization approach has superior noise resistance compared to the PCA approach.

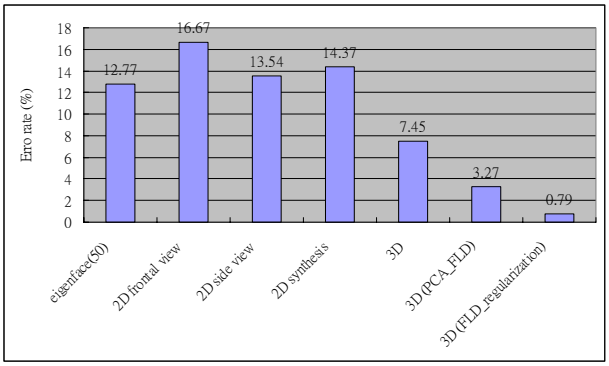

Figure 3. The relative performance of the face recognition algorithms, showing the reduced error rate using a 3D reconstruction and further improvement from FLD and regularization.

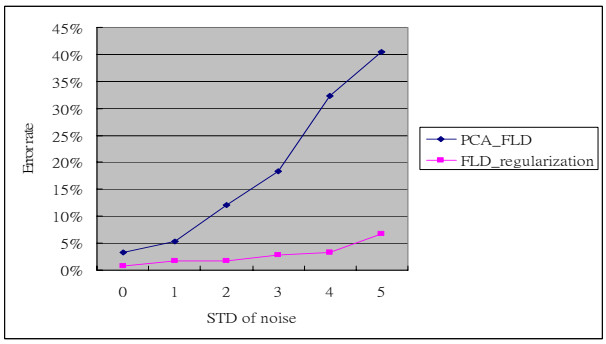

Figure 4. As noise is added to the extracted features, the regularized FLD recognition error remains relatively stable, while the PCA FLD approach shows rapid deterioration.

5. CONCLUSION AND FUTURE WORK

In conclusion, we have demonstrated that 3-dimensional spatial information can provide clear added assistance in face recognition. Furthermore, the regularization of FLD not only improves the performance of face recognition, but also makes the recognition more robust to effects of noisy data.

While we have demonstrated robustness to synthetically degraded features, in the future we plan to investigate how errors associated with automatic extraction of facial features affect the outlined approach, and what level of degradation will still result in acceptable performance.

We have evaluated our approach on a large standard face image database of 68 people with multiples poses from each person. Our next steps will extend this work to the Informedia project's broadcast video collection. The challenge will be to see if the approach scales to the thousands of different human faces that are depicted in broadcast news. This will help us to understand the scalability of FLD as a recognition feature selector.

Finally, because our approach does not use color or texture as features, it is obviously desirable to combine the results of the Eigenfaces method or similar recognition approaches with our approach to further enhance recognition accuracy. Thus we feel our approach of using 3D face reconstruction and regularized Fisher's linear discriminant will not only be effective on its own, but can be utilized to enhance a number of other facial recognition techniques that are already being used.

6. ACKNOWLEDGMENTS

This research is partially supported by the Advanced Research and Development Activity (ARDA) under contract numbers MDA908-00-C-0037.

7. REFERENCES

[1] S. Satoh, Y. Nakamura and T. Kanade, "Name-It: Naming and Detecting Faces in News Video", IEEE Multimedia 6(1), 1999.

[2] M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience vol.3, 1991, pp. 71-86.

[3] T. Kanade, "Picture Processing by Computer Complex and Recognition of Human Faces", PhD thesis, Kyoto University, 1973.

[4] L. Wiskott, J. M. Fellous, N. Kruger and C. V. D. Malsburg, "Face Recognition by Elastic Bunch Graph Matching", IEEE Trans. Pattern Analysis and Machine Intelligence vol.19, July 1999, pp. 775-779.

[5] K. Lander, F. Christie and V. Bruce, "The Role of Dynamic information in the recognition of famous faces", Memory and Cognition, Nov 1999.

[6] H. S. Ip, and L. Yin, "Constructing a 3D individualized head model from two orthogonal views", The visual computer vol. 12, Springer, 1996, pp. 254-266.

[7] T. Sim, S. Baker and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database", Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, May, 2002.

[8] M. Christel, A. G. Hauptmann, and H. Wactlar, "Improving Access to Digital Video Archives through Informedia Technology", Journal of the Audio Engineering Society 49(7/8), July/August 2001, http://www.informedia.cs.cmu.edu